

Reconhecimento de Emoções em Tempo Real utilizando detecção por filtro HaarCascade e Redes Neurais Convolucionais

Ulisses Lopes da Silva
Instituto Tecnológico de Aeronáutica
São José dos Campos/SP - 2025

Resumo—Este trabalho propõe um sistema para reconhecimento de emoções faciais em tempo real a partir de imagens capturadas via webcam, utilizando Redes Neurais Convolucionais (CNNs) para classificação e o classificador HaarCascade para detecção facial. O modelo foi treinado com o dataset FER2013, composto por 35.887 imagens em tons de cinza com resolução de 48×48 pixels, distribuídas em sete categorias emocionais: *angry*, *disgust*, *fear*, *happy*, *sad*, *surprise* e *neutral*. A rede foi projetada com quatro blocos convolucionais seguidos por camadas densas e treinada com o otimizador Adam e a função de perda *categorical crossentropy*. Durante o treinamento, o modelo alcançou uma acurácia de aproximadamente 67,6% nos dados de validação, com estabilidade entre perda e desempenho observada após a 25ª época, em relação às 70 totais. A matriz de confusão revelou desempenho consistente nas classes *happy*, *neutral* e *surprise*, com maior índice de confusão entre *fear*, *sad* e *angry*. O sistema resultante foi capaz de realizar previsões em tempo real com boa generalização para expressões bem definidas em ambientes com boa iluminação. Todavia, os resultados demonstraram que, para uma maior aplicabilidade em interfaces homem-máquina, ambientes educacionais adaptativos e sistemas de monitoramento emocional, seria recomendável refinar o modelo convolucional e diversificar o dataset, bem como aumentar o número de épocas de treinamento e aplicar mais técnicas para evitar *overfitting*.

I. INTRODUÇÃO

Nas últimas décadas, os avanços em inteligência artificial e aprendizado de máquina revolucionaram diversas áreas do conhecimento, permitindo que sistemas computacionais realizem tarefas antes restritas à cognição humana. Dentro das diversas arquiteturas desenvolvidas nesse contexto, as Redes Neurais Convolucionais (CNNs — *Convolutional Neural Networks*) destacam-se por sua eficácia em problemas relacionados ao processamento e análise de imagens.

As CNNs são especialmente projetadas para explorar a estrutura espacial de dados bidimensionais, como imagens, utilizando camadas convolucionais que atuam como filtros detectores de padrões visuais, tais como bordas, texturas e formas. Esta capacidade torna as CNNs ferramentas altamente eficazes em tarefas como classificação de imagens, detecção de objetos, segmentação semântica, diagnóstico médico por imagem, sistemas de vigilância automatizados, veículos autônomos e biometria.

Entre essas aplicações, o reconhecimento de expressões faciais tem ganhado destaque, especialmente por sua relevância em áreas como interação humano-computador, segurança, psicologia, monitoramento comportamental e as-

sistência médica. Reconhecer automaticamente as emoções humanas por meio de expressões faciais pode proporcionar avanços significativos no desenvolvimento de sistemas mais empáticos, personalizados e responsivos.

Este trabalho propõe o desenvolvimento de um sistema capaz de identificar, em tempo real, expressões faciais captadas por uma webcam. Para isso, utiliza-se uma combinação entre técnicas clássicas de detecção facial (filtros HaarCascade) e um modelo baseado em redes neurais convolucionais treinado sobre o dataset FER2013, amplamente utilizado na literatura para problemas de reconhecimento de emoções. O estudo visa avaliar o desempenho do modelo em um cenário prático, explorando os desafios de generalização, ambiguidade das expressões faciais e limitações impostas pelas condições reais de captura de imagem.

II. REVISÃO TEÓRICA

A. Redes Neurais Convolucionais (CNNs)

As Redes Neurais Convolucionais (CNNs) são uma classe de modelos amplamente utilizadas para o reconhecimento e a classificação de imagens. Diferentemente das redes neurais totalmente conectadas, as CNNs são compostas por camadas especializadas que extraem, de maneira hierárquica, características visuais relevantes da entrada. Estas camadas incluem convoluções, funções de ativação, normalização, subamostragem e camadas densas.

A estrutura arquitetônica típica de uma CNN inicia-se com uma ou mais camadas convolucionais, cujos filtros (ou kernels) percorrem a imagem em janelas deslizantes, extraíndo padrões locais. A seguir, aplica-se uma função de ativação não linear — comumente a ReLU ou LeakyReLU — que introduz não linearidade ao modelo. Em seguida, camadas de *pooling* são utilizadas para reduzir a dimensionalidade espacial e a complexidade computacional, preservando as informações mais relevantes. Normalmente, camadas de *batch normalization* também são empregadas para estabilizar e acelerar o treinamento. Após os blocos convolucionais, os dados são achatados e encaminhados a camadas totalmente conectadas (densas), que consolidam as informações extraídas para efetuar a previsão.

Neste trabalho, foi empregada uma arquitetura composta por quatro blocos convolucionais, cada um contendo: uma camada Conv2D, seguida de BatchNormalization, ativação ReLU, e MaxPooling2D, com uso de Dropout para regularização. Após a extração de características, o

modelo conta com três camadas densas, com a última camada utilizando ativação softmax para classificar a imagem em uma das sete categorias emocionais: *angry*, *disgust*, *fear*, *happy*, *sad*, *surprise* e *neutral*.

O treinamento da rede foi conduzido utilizando o algoritmo Adam como otimizador e a função de perda categorical_crossentropy, adequada a problemas de classificação multiclasse. A métrica de avaliação principal foi a accuracy tanto em treino quanto em validação, cujos resultados serão posteriormente discutidos.

B. Filtro HaarCascade para Detecção Facial

Para que o sistema funcione em tempo real a partir da webcam, é necessário detectar a presença de rostos em cada quadro capturado. Para isso, utilizou-se o método de detecção baseado em cascata de classificadores Haar (*HaarCascade*), uma abordagem clássica desenvolvida por (Viola & Jones, 2021).

O filtro HaarCascade baseia-se no uso de características Haar — padrões retangulares que representam variações de intensidade de pixels — que são extraídas da imagem e comparadas com padrões previamente treinados em um conjunto de imagens positivas (com rostos) e negativas (sem rostos). O processo de detecção envolve a varredura da imagem em diferentes escalas e posições por uma janela de busca, verificando se a região analisada contém um rosto.

A cascata de classificadores funciona como um conjunto de filtros encadeados. Cada estágio da cascata aplica um classificador simples e decide se a região da imagem deve ser descartada ou encaminhada ao próximo estágio. Essa abordagem proporciona um alto desempenho com baixo custo computacional, o que a torna ideal para aplicações em tempo real.

Neste projeto, o classificador HaarCascade foi utilizado apenas para localizar e extrair a região da face nos quadros capturados, que posteriormente são redimensionadas para 48×48 e repassadas ao modelo de CNN para classificação emocional.

III. DESENVOLVIMENTO

A. Preparação do Dataset

O dataset utilizado neste projeto é o FER2013 (Facial Expression Recognition 2013), disponibilizado originalmente no Kaggle. Ele contém 35.887 imagens em tons de cinza, com resolução de 48×48 pixels, distribuídas em sete categorias de emoção: *angry*, *disgust*, *fear*, *happy*, *sad*, *surprise* e *neutral*. As imagens foram organizadas em duas pastas principais: uma para treinamento (com 28.709 amostras) e outra para validação (com 7.178 amostras).

Para facilitar o carregamento, as imagens foram estruturadas em subpastas nomeadas com o rótulo de suas respectivas emoções. A biblioteca ImageDataGenerator, do Keras, foi utilizada para realizar o pré-processamento das imagens. Além da normalização dos pixels no intervalo $[0, 1]$, foi aplicada uma estratégia de aumento de dados (data augmentation), incluindo rotações aleatórias, zoom, deslocamentos horizontais e verticais e flips horizontais,

a fim de reduzir o sobreajuste e aumentar a robustez do modelo.

B. Construção da Rede Neural Convolucional

A arquitetura da CNN foi implementada com a API Keras (TensorFlow backend). A rede é composta por quatro blocos convolucionais sucessivos, cada um contendo:

- Uma camada Conv2D com filtro de tamanho 3×3 e ativação ReLU;
- Uma camada de normalização em lotes (BatchNormalization);
- Uma camada de MaxPooling2D;
- Uma camada de Dropout, com taxa de 0.25.

Após os blocos convolucionais, o volume de saída é achatado com uma camada Flatten, seguido por três camadas densas (Dense), com 64, 64 e 7 neurônios, respectivamente, todas com ativação ReLU, normalização em lotes e dropout, exceto a última. A camada final é uma Dense com 7 saídas (correspondendo às emoções) e ativação softmax, adequada para classificação multiclasse.

O modelo foi compilado com a função de perda categorical_crossentropy, otimização via Adam e métrica de acurácia.

C. Aplicação do Filtro HaarCascade

Para realizar a detecção facial em tempo real via webcam, foi utilizado o classificador HaarCascade da biblioteca OpenCV. Esse método é baseado em um algoritmo de aprendizado supervisionado em cascata, com uso de retângulos de Haar para extração de características.

Durante a execução do sistema, o vídeo capturado pela webcam é processado quadro a quadro. Em cada frame, o filtro HaarCascade é aplicado para localizar regiões faciais. As faces detectadas são extraídas, redimensionadas para 48×48 pixels e normalizadas antes de serem passadas à CNN treinada para inferência. O resultado da classificação é sobreposto ao vídeo em tempo real, mostrando a emoção detectada.

D. Etapas de Treinamento

O treinamento foi realizado por 70 épocas, com batch_size igual a 32. Ao longo das épocas, observou-se uma melhora consistente na acurácia até aproximadamente a 25ª época, após a qual o desempenho se estabilizou. A função de perda diminuiu progressivamente e não indicou sinais de sobreajuste evidente, graças à aplicação de técnicas como normalização em lotes, dropout e aumento de dados.

Ao final, o modelo obtido foi salvo em disco e utilizado para realizar inferências ao vivo em vídeos capturados pela webcam, possibilitando uma aplicação funcional de reconhecimento de emoções faciais em tempo real.

IV. RESULTADOS

A. Desempenho Durante o Treinamento

Durante as 70 épocas de treinamento, o modelo demonstrou evolução razoável tanto na acurácia quanto na minimização da função de perda. A acurácia de validação

atingiu um valor final de aproximadamente 67,6%, conforme observado no gráfico da Figura 1 e no arquivo de métricas finais. A função de perda de validação, por sua vez, estabilizou-se em torno de 1.03, sugerindo convergência sem indícios evidentes de sobreajuste.

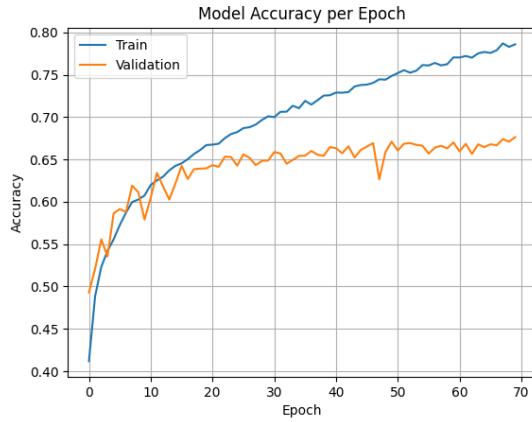


Fig. 1. Acurácia por época no treinamento e validação.

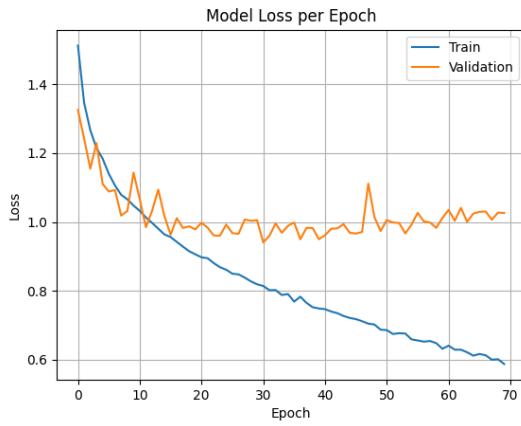


Fig. 2. Perda por época no treinamento e validação.

B. Avaliação por Classe

O relatório de classificação (Tabela I) evidencia um desempenho heterogêneo entre as classes. Emoções como *happy*, *neutral* e *sad* apresentaram valores altos de **precision**, **recall** e **F1-score**. Isso se deve, em parte, à sua forte representação no conjunto de treinamento: *happy* (7215 imagens), *neutral* (4965 imagens) e *sad* (4830 imagens). Esse número elevado de amostras fornece ao modelo maior variabilidade e representatividade, facilitando o aprendizado de padrões distintos para essas emoções.

Em contrapartida, a classe *disgust*, com só 436 imagens no treino e 111 na validação, apesar de apresentar valores razoáveis em **precision**, **recall** e **F1-score** em comparação com as demais, obteve desempenho ruim na simulação em tempo real, sugerindo que o fato de haver poucas amostras

prejudicou a capacidade de generalização do modelo. Dessa forma, muitas vezes a expressão de *disgust* era confundida com *angry* ou *sad* na simulação. Isso evidencia o impacto do desequilíbrio de classes na performance do classificador.

TABLE I
RELATÓRIO DE CLASSIFICAÇÃO POR CLASSE (VALIDAÇÃO)

Classe	Precision	Recall	F1-score	Suporte
Angry	0.61	0.59	0.60	958
Disgust	0.74	0.63	0.68	111
Fear	0.55	0.47	0.51	1024
Happy	0.89	0.86	0.87	1774
Neutral	0.57	0.72	0.64	1233
Sad	0.59	0.51	0.55	1247
Surprise	0.72	0.83	0.77	831
Acurácia			0.68	7178
Média Macro	0.67	0.66	0.66	7178
Média Ponderada	0.68	0.68	0.67	7178

C. Matriz de Confusão

A matriz de confusão (Figura 3) mostra que a maioria das classificações incorretas ocorre entre pares de emoções com expressões faciais similares, como *fear* e *surprise* ou *sad* e *neutral*. A classe *disgust*, devido à sua baixa representação, foi frequentemente confundida com *angry* e *sad*, indicando dificuldade do modelo em capturar seus padrões visuais únicos. Como muitas confusões ocorrem entre emoções visualmente próximas, isso sugere que, mesmo com boa acurácia global em relação à literatura, o modelo ainda carece de refinamento para separar nuances sutis de expressões faciais.

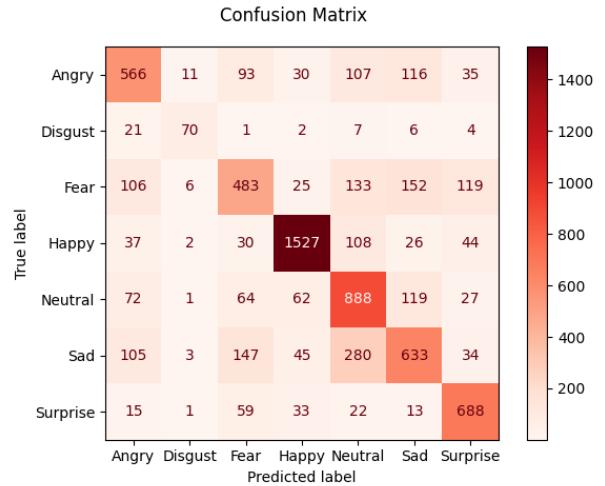


Fig. 3. Matriz de confusão da predição sobre o conjunto de validação.

D. Simulações em tempo real

Para validar o funcionamento prático do modelo treinado, foi realizada uma simulação em tempo real utilizando a webcam do computador. Durante essa etapa, rostos foram detectados em tempo real por meio do classificador Haar-Cascade, e as emoções foram inferidas a partir do modelo convolucional previamente treinado.

As Figuras 4 a 10 ilustram alguns quadros capturados da simulação, nos quais é possível observar a sobreposição do rótulo da emoção prevista sobre a região do rosto detectado.

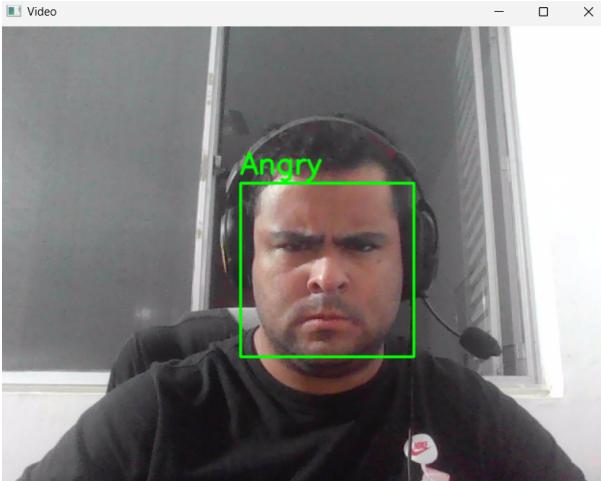


Fig. 4. Predição correta da emoção “Angry”.

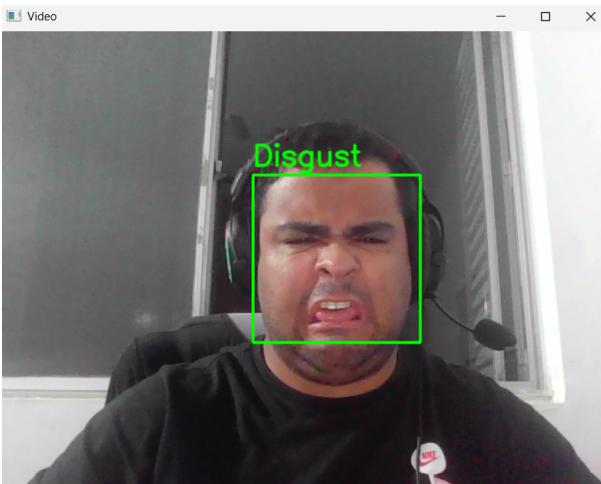


Fig. 5. Predição correta da emoção “Disgusting”.

Os resultados visuais reforçam a capacidade do sistema em identificar corretamente diferentes expressões faciais em tempo real, confirmando a razoável eficácia da solução desenvolvida para aplicações práticas. Não obstante, o modelo está plenamente apto a receber melhorias futuras.

V. DISCUSSÃO DOS RESULTADOS

A. Causas Prováveis de Desempenho Heterogêneo

O desempenho significativamente superior em classes como *happy* e *neutral* está diretamente relacionado à sua maior representatividade no dataset. Além disso, essas emoções tendem a apresentar expressões faciais mais distintas, o que favorece a sua identificação por redes convolucionais. Já classes como *disgust*, além de escassas, apresentam expressões que podem ser confundidas com *angry* ou *sad*, o que reforça a importância do balanceamento

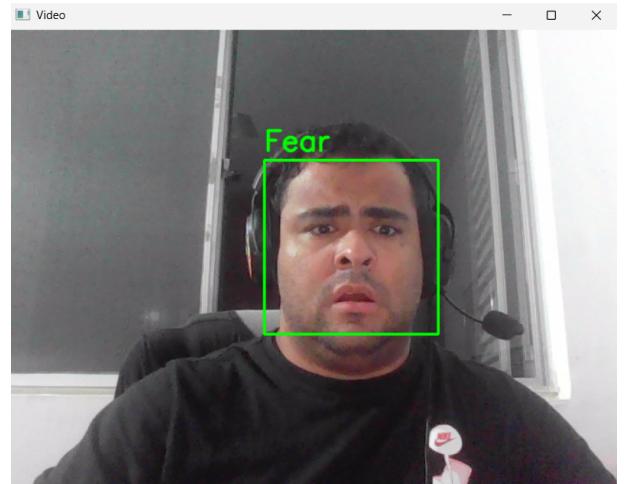


Fig. 6. Predição correta da emoção “Fear”.

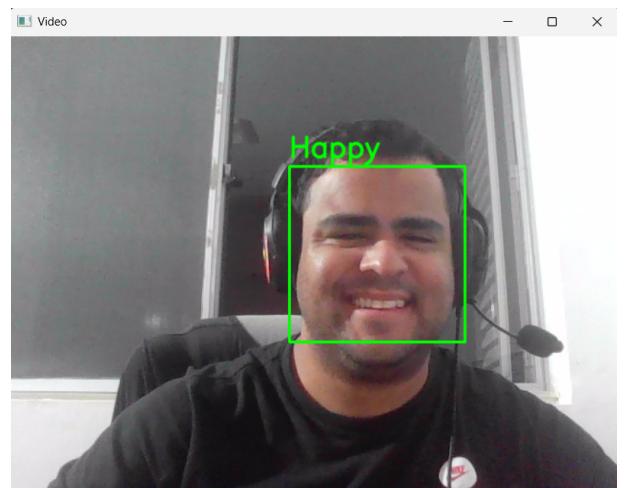


Fig. 7. Predição correta da emoção “Happy”.

de classes. Não obstante, o valores relativamente bons das métricas no quadro de classificação, especialmente os da classe *neutral*, em comparação com outras classes, mascara o fato de que, com poucas amostras, a capacidade de generalização do modelo é prejudicada, diminuindo sua performance em tempo real.

B. Propostas de Melhoria

Com base nos resultados observados, algumas melhorias podem ser propostas para aprimorar o desempenho do sistema:

- **Balanceamento do dataset:** O uso de técnicas como *oversampling* das classes minoritárias (e.g., *disgust*) ou *class weights* no treinamento pode ajudar a reduzir o viés do modelo. Além disso, diversificá-lo com imagens de *databases* como *AffectNet* e *CK+* podem ajudar na capacidade de generalização do modelo.
- **Aumento da profundidade da rede:** Arquiteturas mais profundas como *ResNet* ou *MobileNet*, combinadas com regularização adequada, podem capturar padrões mais

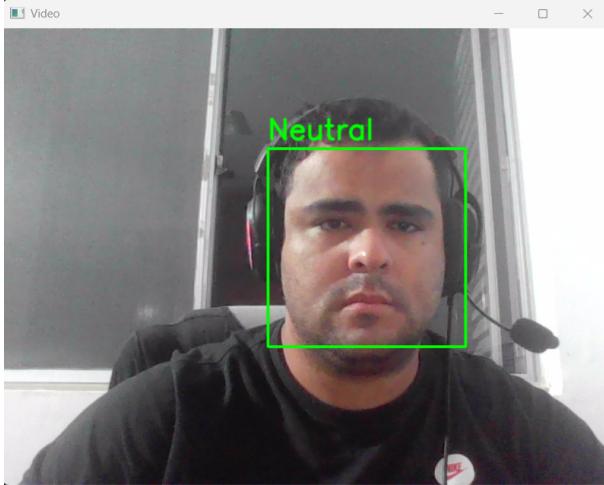


Fig. 8. Predição correta da emoção “Neutral”.

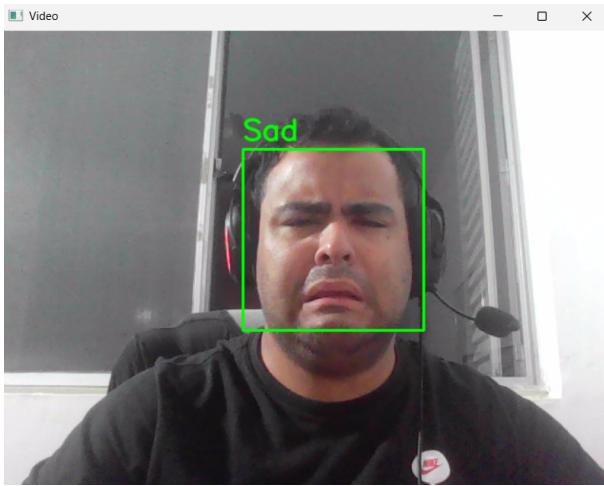


Fig. 9. Predição correta da emoção “Sad”.

refinados, aumentando a acurácia.

- **Uso de transfer learning:** Inicializar a rede com pesos pré-treinados em bancos maiores (como o ImageNet) pode acelerar a convergência e melhorar a generalização.
- **Pré-processamento com detecção de face mais robusta:** Substituir o HaarCascade por detectores mais precisos, como MTCNN ou o MediaPipe FaceMesh, pode melhorar a qualidade das regiões extraídas para inferência.
- **Cross-validation:** Adotar validação cruzada estratificada pode fornecer estimativas mais confiáveis do desempenho real do modelo.
- **Data augmentation direcionado:** Aumentar a variabilidade visual das classes minoritárias por transformações específicas (zoom, rotação, ruído) pode ajudar no aprendizado.

Estas ações podem ser incorporadas em estudos futuros, visando o desenvolvimento de um sistema de detecção emocional mais robusto e generalizável para aplicações em

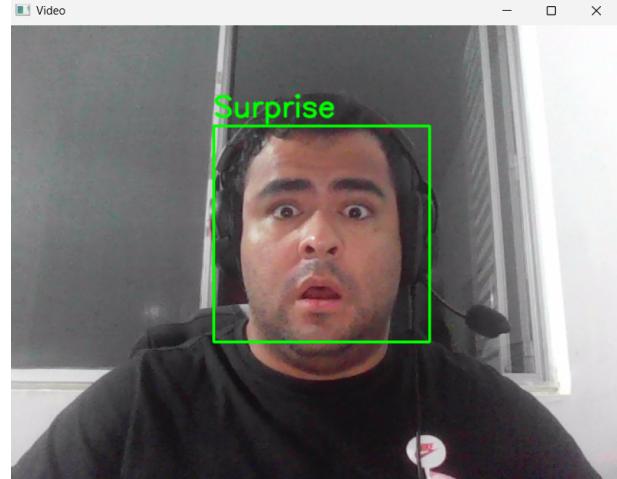


Fig. 10. Predição correta da emoção “Surprise”.

ambientes reais.

VI. CONCLUSÃO

Este trabalho apresentou o desenvolvimento de um sistema para reconhecimento de emoções faciais em tempo real utilizando uma rede neural convolucional (CNN) treinada sobre o dataset FER2013, com suporte à detecção de rostos por meio do filtro HaarCascade. Os resultados demonstraram desempenho consistente, com uma acurácia global com cerca de 68% na validação e f1-scores superiores a 0.75 para classes como *happy* e *surprise*.

A análise revelou que classes com maior representatividade no conjunto de dados tendem a apresentar melhores métricas e melhor desempenho na simulação em tempo real, enquanto categorias com poucas amostras, como *disgust*, apesar de métricas razoáveis, são mais suscetíveis a erros de classificação no ambiente de simulação. Esse comportamento ressalta a importância do balanceamento de dados no desempenho de classificadores supervisionados, bem como ressalta a tanto a complexidade quanto a riqueza do estudo de CNNs para esse tipo de aplicação, dada a diversidade de nuances possíveis de serem percebidas no rosto de cada indivíduo.

Como proposta de trabalhos futuros, destaca-se a adoção de técnicas de aumento de dados, arquiteturas de redes mais profundas (como ResNet ou EfficientNet) e a aplicação de métodos de atenção para reforçar a extração de características faciais relevantes. A integração com múltiplos detectores de face, como o MediaPipe, também pode ser considerada para maior robustez em cenários variados de iluminação e orientação facial.

REPOSITÓRIOS

Os códigos-fonte desenvolvidos, incluindo as implementações da rede neural, scripts de treinamento e simulação, bem como também as imagens e arquivos de texto relativos aos testes e validações realizadas, estão disponíveis no repositório público:

- <https://github.com/ulissesuls/Real-Time-Emotions-Detection>

REFERÊNCIAS

- [1] Goodfellow et. al., "Challenges in Representation Learning: A report on three machine learning contests", Unicer, 2013. Disponível em: <http://arxiv.org/abs/1307.0414>
- [2] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge, Massachusetts (2016). Disponível em: <http://www.deeplearningbook.org>
- [3] D. P. Lopez, F. Z. Canal, G. G. Scotton, E. Pozzebon e A. C. Sobieranski, "Uma abordagem computacional em tempo real para reconhecimento de expressão facial humana baseada na extração de recursos de referência" *SEVENED – Revista de Saúde, Educação e Meio Ambiente*, vol. 4, n. 1, pp. 4–6, 2024. Disponível em: <https://doi.org/10.56238/sevened2024.004-006>
- [4] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)", *SN Applied Sciences*, vol. 2, no. 3, pp. 1–8, Feb. 2020. DOI: <https://doi.org/10.1007/s42452-020-2234-1>.
- [5] DESHPANDE, A. "The 9 deep learning papers you need to know about (understanding cnns part 3)". University of California, Los Angeles (UCLA),
- [6] R. Ratan, "Building an Emotion Detector with LittleVGG", GitHub repository, [Online]. Available: <https://github.com/rajeevratan84/DeepLearningCV/blob/master/18.2>
- [7] F. Z. Canal, "Método para Reconhecimento em Tempo Real de Expressões Faciais em Grupos utilizando Redes Neurais Convolucionais", Dissertação de Mestrado, Universidade Federal de Santa Catarina, Programa de Pós-Graduação em Tecnologias da Informação e Comunicação, Araranguá, 2024.